Center for European Studies
Working Paper Series 131

The Socioeconomic Diversity of European Regions

by

Cristina del Campo
Dpto de Estadística e I.O. II
Universidad Complutense de Madrid
Campus de Somosaguas, 28223
Madrid, Spain
campocc@ccee.ucm.es

Carlos M. F. Monteiro
João Oliveira Soares
CEG-IST, Instituto Superior Técnico,
Universidade Técnica de Lisboa
1049-001 Lisbon, Portugal
cmont@mail.telepac.pt
joaosoares@ist.utl.pt

Abstract

It is well-known that there are significant differences among the European Union regions, which have been heightened due to the most recent enlargement in 2004. This paper aims to analyze this diversity and propose a classification of European Regions (EU) that is adjusted to the different axes of socioeconomic development and, simultaneously, is useful for European regional policy purposes.
The data used in this paper were published by the European Union Statistical Office (Eurostat) and correspond to the main statistical indicators of NUTS2 (Nomenclature of Territorial Units for Statistics) regions in the EU. Multivariate statistical techniques allowed the identification of clusters of socioeconomic similarity, which are contrasted with the classes considered in the financial proposal of the European Commission (EC) for the period 2007-2013. It was found that each of the two main groups of the EC classification – *convergence regions* and *competitiveness and employment regions* – comprises at least two significantly different groups of regions, which differ not only in their average income but also in other indicators associated with their particular weaknesses. Also, it has been revealed that two other groups–*phasing-in regions* and *phasing-out regions* –, beyond their inexpressive denomination, lack homogeneity, being spread throughout different clusters.

Keywords: regional development; regional indicators; multivariate statistical analysis; cluster analysis; factor analysis; European policy.

JEL Classification:; R1; R58; O52; O1; C1

1. Introduction

The roots of the European Union go back to the post-World War II period, when the French Foreign Affairs Minister Robert Schuman proposed the creation of a European Coal and Steel Community, based on an original idea by Jean Monnet. Through the years this Community evolved: first, to the European Economic Community (EEC), founded in 1957 by the Treaty of Rome; later, to the present European Union (EU), born in 1993 after the Maastricht Treaty. During this economic and political integration process, various steps led to the successive inclusion of more nations. Currently, with the adhesion of ten new countries on May 1, 2004, the enlarged Union includes 454 million inhabitants living in twenty-five countries. Two more – Romania and Bulgaria – are expected to become members in 2007, and others, like Turkey, expect to join the EU in the future.

These successive enlargements, and particularly the last one, have heightened regional disparities within the EU. "Average per capita income in the EU of Twenty-Five will be 12.5 percent less. The economic and social disparities will double … 18 percent of the Community … continues to represent half of its wealth and three-quarters of research capability! We will not have sustainable growth with a countryside that is empty and cities that are choking!" These are the words of the member of the European Commission responsible for regional policy and institutional reform (Barnier, 2004).

In order to attack these asymmetries and fight for cohesion, the European Commission established several priorities for 2007-2013. The first one concerns the *convergence* of least developed countries and regions, involving those regions in the EU whose per capita GDP is less than 75 percent of the Community's average, the older Objective 1 (thirty-three *convergence regions* from EU15 and 37 from the new Member States. See Figure 1). The second priority is regional *competitiveness and employment*, focused on the sustainable growth problems of the more developed regions (in *latus sensus*, meaning more than 75 percent of the Community's average), involving policies constructed around the innovation and knowledge economy. Furthermore, the Commission has distinguished two other groups. One, the *phasing-in regions* (twelve regions, eleven from the EU-15 and one Hungarian region. See Figure 1). They comprise those regions which recently came out of Objective 1 and, so, will have easier access to funds allocated under the competitiveness objective. The other group, comprising sixteen EU-15 regions, corresponds to those regions which would have continued to belong to Objective 1 if they had not suffered from the statistical effect of the enlargement to twenty-five countries (decrease of Community's average GDP per head). Those *phasing-out regions* will temporarily have preferential financial treatment.

It is certainly possible to draw a picture of the European regions based exclusively on their GDP figures. However, this is an incomplete and static picture – it does not account for the potential development prospects associated, for example, with population density, demographic distribution, or education/qualification of labor. Thus, although it is true that the recent European enlargement increased enormously the disparity of income inside the EU, it is also certain that many of the less developed regions that recently joined the Union possess characteristics of competitiveness – youth and education – not found in the more depressed, abandoned, rural regions of former EU members. As the Commission points out, "Regions with problems of competitiveness … are not confined to the Cohesion countries in the present EU and the new Member States" (European Commission; 2004).

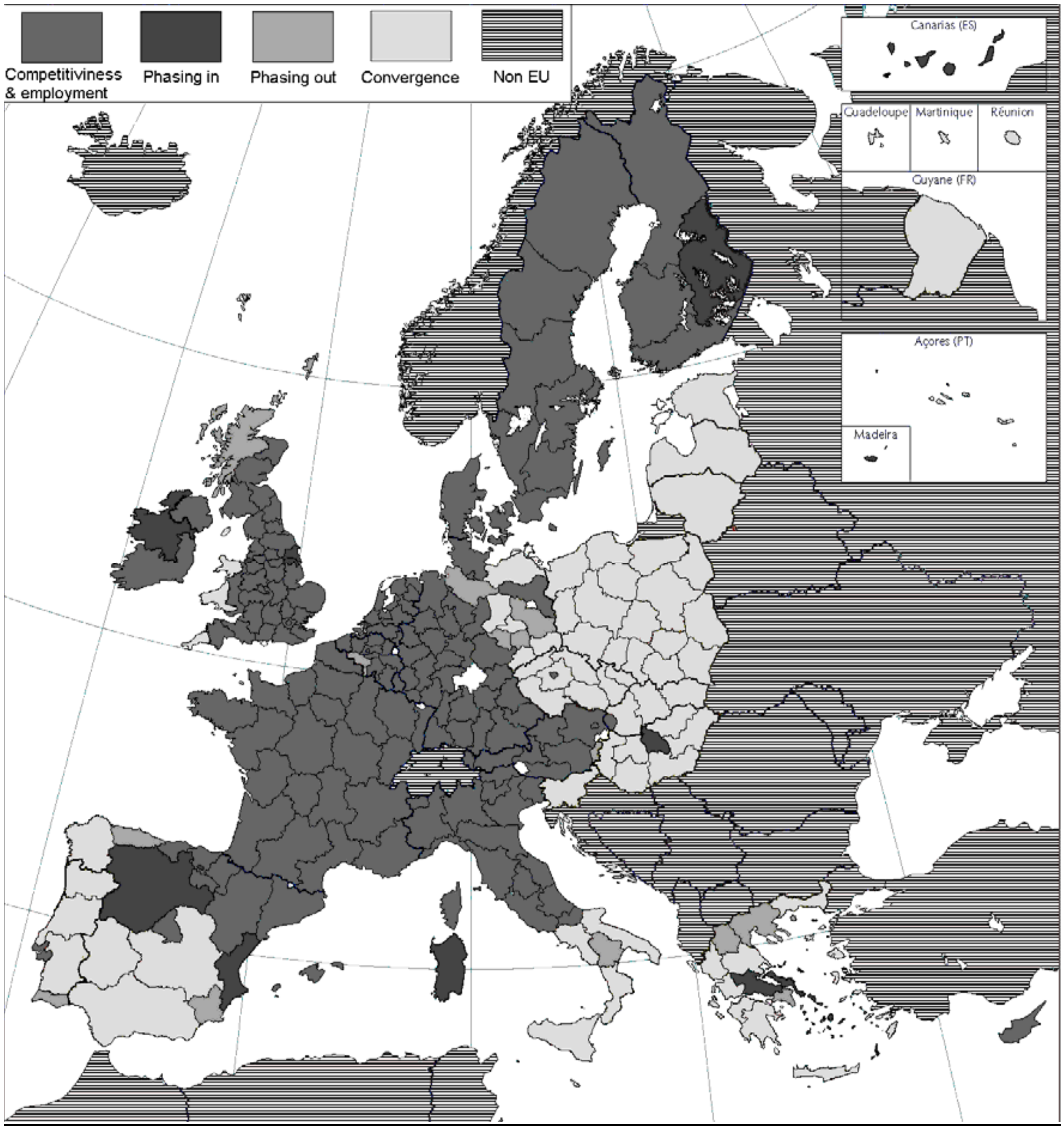


Figure 1 – EU25: Convergence and Competitiveness Objectives 2007-2013 (based on Eurostat GDP/head data available on 04/04/2005)

How, then, can we draw a global picture that confronts this reality with the aforementioned GDP-based classification, which supports European regional policy and inherent allocation of financial resources? The answer to this question constitutes the main goal of this paper. Simultaneously, we believe that the methodological approach that we followed may motivate other authors to discuss it and present different contributions in the future.

In order to achieve the previously mentioned goal, data published by the European Union Statistical Office (Eurostat) will be used. The Eurostat data are released annually in different levels, called NUTS (Nomenclature of Territorial Units for Statistics). NUTS2 regions are an adequate unit of analysis since they correspond to the administrative structure of the Member States' major regional areas (Länder in Germany, Régions in France, Comunidades Autónomas in Spain, Regioni in Italy, etc.) and, hence, may enable the authors to establish disparities, if any, even inside each country. In addition, NUTS2 regions are the basis for allocation of financial resources in the context of economic and social cohesion policy. Therefore, socioeconomic variables measured in NUTS2 regions were chosen for further analysis using multi-variate statistical techniques.

Other papers applying multi-variate statistical analysis to socioeconomic problems and, particularly, to the classification of different types of administrative divisions (municipalities, counties or regions) can be found in the literature: Cziráky *et al.* (2005), Aragon *et al.* (2003), González and Morini (2000), Soares *et al.* (2003), Peschel (1998), Pettersson (2001), Rovan and Sambt (2003) or Rúa Vieites *et al.* (2003). Those studies are restricted to a smaller area inside Europe, specifically Croatia, the Midi-Pyrénées Region, Tenerife Island, Portugal, the Baltic Sea countries, a Swedish county, Slovenia and the Spanish region of Galicia in their respective cases. There are other contributions outside of Europe. For example, Stimson *et al.* (2001) focused on the United States of America and Hill *et al.* (1998) on Australia. We did find one study on the EU15, by Rúa Vieytes *et al.* (2000), but it has a different focus and lacks data on many regions. Besides, most of these data are from 1994.

The rest of the paper is organized as follows: in Section 2, we describe and characterize the data used in this work. Section 3 is focused on reducing the information accounted for in all the Eurostat variables to a few socioeconomic dimensions, based on factor analysis. Section 4 is dedicated to finding clusters of regions presenting similar characteristics of development. Conclusions are stated in Section 5.

2. Data description

The socioeconomic variables considered in the present study (Table 1) were selected from the Eurostat Database "Regio." They correspond to nineteen of the twenty-four Main Regional Indicators published in the *Third report on economic and social cohesion* (European Commission, 2004). The five variables that were excluded did not satisfy the requirements that were established for proceeding with the analysis – i.e., up-to-date information, availability for all the twenty-five EU countries, and expression as ratios, in order to avoid scale problems. Since data for the regions whose frontiers changed in May 2003* were not available, the following decisions concerning missing data were made. Where changes were small and "old" data were available, these data were retained; otherwise, the existence of missing points was assumed and, consequently, those regions were not integrated in the subsequent analysis (thirteen of the 254 existing NUTS2 regions were left out of the analysis).

Table 1 – Regional Indicators Considered in the Study

Code	Description	Year
Demography		
popdens	population density (inh./km²)	2003
pop014	percentage of the population aged less than 15 years	2003
pop1564	percentage of population between 15 and 64 years	2003
pop65	percentage of the population aged more than 65	2003
Economy		
gdppps	GDP/head (PPS), EU25=100	2002
empagr	agriculture employment (% of total)	2002
empind	industry employment (% of total)	2002
empserv	services employment (% of total)	2002
patent	EPO patent applications per million inh., average	1999-2003
Employment		
emptot	total employment rate (ages 15-64 as % of pop. ages 15-64)	2003
empf	female employment rate (ages 15-64 as % of pop. ages 15-64)	2003
empm	male employment rate (ages 15-64 as % of pop. ages15-64)	2003
unemptot	total unemployment rate (%)	2003
unemplt	long term unemployed(% of total unempl.)	2003
unempf	female unemployment rate (%)	2003
unempy	youth unemployment rate (%)	2033
Education		
lowedu	percentage of active population with pre-primary, primary and lower secondary education (levels 0-2 ISCED 1997)	2003
mededu	percentage of active population with upper secondary and post-secondary non-tertiary education (levels 3-4 ISCED 1997)	2003
highedu	percentage of active population with tertiary education (levels 5-6 ISCED 1997)	2003

The descriptive statistics shown in Table 2 reflect some huge asymmetries between the EU regions. The most remarkable ones are population density (a 1:4500 ratio between the lowest and the highest densities) and patents (0:745). As for demography, the differences in potentially active population (47.6 percent:72.6 percent) or in the percentage of aged population (8.5 percent:31 percent) are also significant. In turn, GDP per capita shows a dispersion of 1:10, unemployment a dispersion of 1:15, which turns out to be 1:20 if only long-term unemployment is considered. This is also the ratio found by looking at the percentage of active population with tertiary education. Finally, it should be noted that some of the variables show excess kurtosis or skewness and, therefore, do not follow normal distributions, a fact that was taken into account when choosing the techniques to be used in the following sections.

Table 2 – Descriptive Statistics of the Regional Indicators

	Min	Max	Mean	St. Dev.	Skew	Kurtosis
popdens	2.10	8946.45	385.78	926.37	5.73	39.62
pop014	10.14	23.33	16.68	2.47	-0.35	-0.06
pop1564	47.60	72.57	66.68	2.68	-1.90	12.59
pop65	8.59	31.10	16.64	3.13	0.70	2.38
gdppps	32.00	315.40	95.52	34.97	1.38	6.50
empagr	0.06	39.40	6.54	7.49	2.36	5.69
empserv	41.04	88.46	65.19	9.71	-0.14	-0.37
empind	11.43	46.33	28.27	7.38	0.04	-0.31
patent	0.00	745.88	106.35	130.31	2.17	5.85
emptot	40.10	78.60	63.35	8.20	-0.50	-0.12
empf	24.00	76.10	55.44	10.29	-0.64	0.18
empm	46.60	85.60	71.24	7.69	-0.77	0.31
unemptot	2.00	31.80	8.95	5.69	1.38	1.42
unemplt	4.09	80.44	38.45	15.80	0.23	-0.77
unempf	2.30	33.30	10.00	6.79	1.24	0.85
unempy	4.20	58.40	18.91	11.96	1.28	1.06
lowedu	3.30	86.30	25.97	17.88	1.12	0.76
mededu	7.13	81.45	49.38	16.36	-0.45	-0.15
highedu	4.80	48.24	23.35	8.20	0.29	-0.24

The correlation between each pair of variables was also computed (Table 3). The most relevant correlations, above 0.5, are shown in bold. Beyond the obvious strong correlations between variables of the same category (demography, economy, employment and education), two aspects can be emphasized. One is the relevant correlation between GDP per capita and, respectively, the weight of services employment (0.62), patents per million inhabitants (0.51), total employment rate (0.50) and high level of education (0.49). The other aspect is the significant positive correlation between high educational level and the weight of services employment (0.58). On the contrary, the correlations between high education level and the weights of industry and agricultural sectors are both negative (-0.31 and -0.44 respectively). Normally, this implies that these sectors generate less value added than services and, consequently, regions where these sectors are more represented, will show less GDP per capita.

3. Regional indicators and socioeconomic dimensions

Socioeconomic similarities among NUTS2 regions can be investigated using the original Eurostat variables. However, in situations where groups of observations are formed using the measured variables, the researcher has to intervene in order to choose which original variables to use (Hair et al., 1998). This choice is decisive in order, for example, to avoid group solutions that could be biased towards "over-measured" characteristics, e.g., characteristics that are represented by more original variables than the others. This happens with the present data, where a different number of regional indicators represent each category – demography, economy, employment and education (Table 1).

Table 3 – Correlation Matrix

	popdens	pop014	pop1564	pop65	gdppps	empagr	empserv	empind	patent	emptot	empf	empm	unemptot	unemplt	unempf	unempy	lowedu	mededu	highedu
popdens	1																		
pop014	0.03	1																	
pop1564	0.18	-0.25	1																
pop65	-0.18	**-0.56**	**-0.66**	1															
gdppps	**0.55**	0.00	-0.05	0.04	1														
empagr	-0.24	-0.06	0.07	-0.01	-0.52	1													
empserv	0.41	0.20	-0.22	0.04	**0.62**	**-0.67**	1												
empind	-0.30	-0.20	0.22	-0.04	-0.27	-0.17	**-0.62**	1											
patent	0.08	0.11	-0.07	-0.02	**0.51**	-0.41	0.28	0.06	1										
emptot	0.04	0.21	-0.27	0.07	**0.50**	-0.46	0.40	-0.05	0.47	1									
empf	0.07	0.26	-0.24	0.01	0.42	-0.46	0.40	-0.05	0.48	**0.93**	1								
empm	0.00	0.11	-0.26	0.13	**0.50**	-0.34	0.31	-0.05	0.35	**0.87**	**0.64**	1							
unemptot	0.00	-0.13	0.32	-0.18	-0.49	0.38	-0.32	0.02	-0.35	**-0.80**	**-0.63**	**-0.85**	1						
unemplt	-0.04	-0.33	0.28	0.01	-0.41	0.37	-0.40	0.15	-0.27	**-0.75**	**-0.66**	**-0.70**	**0.69**	1					
unempf	-0.08	-0.23	0.30	-0.08	-0.49	0.46	-0.37	0.01	-0.43	**-0.85**	**-0.78**	**-0.75**	**0.94**	**0.69**	1				
unempy	-0.04	-0.03	0.15	-0.11	-0.47	**0.51**	-0.29	-0.15	-0.47	**-0.84**	**-0.75**	**-0.77**	**0.84**	**0.62**	**0.88**	1			
lowedu	-0.12	-0.20	-0.06	0.20	-0.04	0.33	-0.16	-0.13	-0.30	-0.26	-0.47	0.08	-0.04	0.06	0.20	0.19	1		
mededu	-0.08	0.16	0.15	-0.25	-0.24	-0.10	-0.15	0.30	0.10	0.06	0.27	-0.24	0.16	0.12	-0.05	-0.03	**-0.86**	1	
highedu	0.34	0.09	-0.12	0.04	0.49	-0.44	**0.58**	-0.31	0.36	0.36	0.39	0.23	-0.17	-0.31	-0.24	-0.26	-0.34	-0.17	1

An approach to dealing with this problem is to use a method of data reduction such as exploratory factor analysis, which is capable of identifying a smaller set of uncorrelated variables. Each of these factors is associated with a set of highly correlated original variables. These derived factors can subsequently be used as the basis for group formation,** with the additional advantage of revealing the underlying structure of the data. In this study, principal components analysis has been used for factor extraction. This common procedure does not make any distribution assumption for the original data and simultaneously enables that a few principal components account for a major proportion of total variance.

For implementing this approach, the first step is the visual examination of the correlation matrix (Table 3). This examination reveals considerable amount of correlation in the data, with all variables having at least one correlation coefficient greater than 0.5. In addition, the determinant of the correlation matrix is null, which further supports the appropriateness of proceeding with factor analysis (Lattin, Carrol and Green, 2003, p.110). For this same reason, the Bartlett (1950) test of sphericity could not be computed.

The next step, the decision on the number of factors to retain, was based on the eigenvalue criterion (Kaiser 1960). Therefore, the first five factors, with eigenvalues greater than 1, were retained (Table 4). The Ludlow (1999) criterion points to the same direction since there is a clear variance diminution after the fifth factor. Moreover, this five-factor solution explains more than 80 percent of the total variance of the original variables, a good match according to Hair et al. (1998). The five-factor structure also gave the best interpretative solution when compared with three, four and six varimax rotated factor structures. This is a relevant criterion since "in practice the researcher is interested in the interpretability and operational significance of the factor solutions" (Lattin, Carrol and Green, 2003).

Table 4 –Principal Components Analysis - Explained Variance

Factor	Eigenvalue	% variance	Cumulative % variance
1	7.23	38.07	38.07
2	2.57	13.53	51.60
3	2.43	12.77	64.37
4	1.56	8.22	72.59
5	1.47	7.72	80.31
6	0.79	4.17	84.47
7	0.71	3.76	88.24
8	0.65	3.40	91.64
9	0.51	2.68	94.32
10	0.33	1.72	96.04
11	0.26	1.37	97.40
12	0.22	1.13	98.54
13	0.16	0.84	99.38
14	0.10	0.54	99.92
15	0.01	0.06	99.98
16	0.00	0.01	100.00
17	0.00	0.00	100.00
18	0.00	0.00	100.00
19	0.00	0.00	100.00

The five-factor solution has three additional merits. Firstly, almost all variables are highly correlated with only one factor. Secondly, all variables have at least one factor loading greater in absolute value than 0.5, which is considered to be very significant (Hair et al., 1998). Lastly, Table 4 shows that this factor structure explains between 62 percent and 98 percent of the variance of each original variable, except for the variable "patent" where it explains only 44 percent of its variance. The derived rotated 5-factor structure is shown in Table 5, with the omission of factor loadings that are smaller in absolute value than 0.45.

Concerning the interpretation of the factors, Table 5 shows that the first three factors are essentially related to three categories of indicators – employment, economy and education. Factor 1, **(Un)employment Factor,** expresses high levels of unemployment, with strong positive correlations with unemployment indicators and negative correlations with employment variables. It can be also noted a (relatively low) negative correlation with the number of patents per million inhabitants, which is an expected result. Factor 2, **Economic Factor**, associated with high levels of GDP per capita and large number of jobs in the service sector, is also related positively to one demographic and one education variable – respectively population density and percentage of active population with tertiary education. Therefore, a region with a high score on this factor is certainly rich, with a wide offer of services, and a modern and essentially urban economy. Factor 3, **Education Factor**, expresses high percentage of active populations with upper secondary and post-secondary levels, and consequently low percentage of active populations with pre-primary, primary and lower secondary education

Table 5 -Varimax Rotated Matrix

	F1	F2	F3	F4	F5	Communalities
popdens		0.71				0.62
pop014					-0.88	0.84
pop1564				0.89		0.88
pop65				-0.82	0.52	0.98
gdppps		0.74				0.81
empagr		-0.54				0.63
empind		-0.54			0.52	0.77
empserv		0.83				0.84
patent	-0.46					0.44
emptot	-0.92					0.92
empf	-0.78					0.85
empm	-0.90					0.86
unemptot	0.93					0.89
unemplt	0.77					0.72
unempf	0.92					0.88
unempy	0.91					0.88
lowedu			-0.95			0.95
mededu			0.91			0.89
highedu		0.74				0.62

As for the other additional factors, the fourth one is associated with a high percentage of active adults ("pop1564") along with a reduced percentage of retired people ("pop65"), and the fifth is highly negatively correlated with the percentage of children in the population ("pop014") and shows a reasonable positive correlation with the percentage of aged people ("pop65"). These two factors are both related with demography – the fourth category of regional indicators – and so, the preservation of both for cluster analysis would lead to the already mentioned "overweight" effect. Since it did not seem desirable to favor demography aspects in the following steps of the study and, at the same time, the interpretability of an imposed 4-factor solution was less obvious, the decision was made to proceed with cluster analysis maintaining only the fourth factor as the **Demography Factor** and leaving out the fifth factor.

4. Multi-dimensionality of economic development and regional clusters

To search for groups of NUTS2 regions different agglomerative hierarchical clustering procedures were carried out, involving the scores of the four factors referred to in Section 3. The objective of this first step was to analyse the agglomeration schedules and dendrograms in order to establish the number of clusters to choose. A dendogram is a two-dimension diagram that illustrates the fusions made at each successive stage of the process. The observations (in this case, the regions) are listed on the horizontal axis and the vertical axis represents the successive steps. The best interpretative cluster solution can be illustrated by the dendrogram shown in figure 2, corresponding to Ward's method and squared Euclidean distances (other authors emphasize the performance of this method – Everitt, 1993, 2001; Punj and Stewart, 1983; Millingan, 1980).

An important problem is how to select the number of clusters. The distances between clusters at successive steps may serve as guide. From the analysis of this dendrogram, namely from the analysis of the successive increases in the distances at which clusters were joined, it can be concluded that a reasonable choice must fall within the three to five clusters range of solutions. Further investigation, starting from the five-cluster solution, revealed that the left-hand cluster in Figure 2 corresponds roughly to *Convergence Regions* of the new eastern Member States. This is the last cluster to merge, as a result of huge differences with the great majority of the EU-15 regions. At the same level, observing Figure 2 from left to right, the second and third clusters comprise a great number of regions from the South of the EU. These clusters are the first ones to merge when evolving to four clusters. Finally, the last two clusters correspond roughly to the two richest regions within the Union. They merge when three clusters are formed.

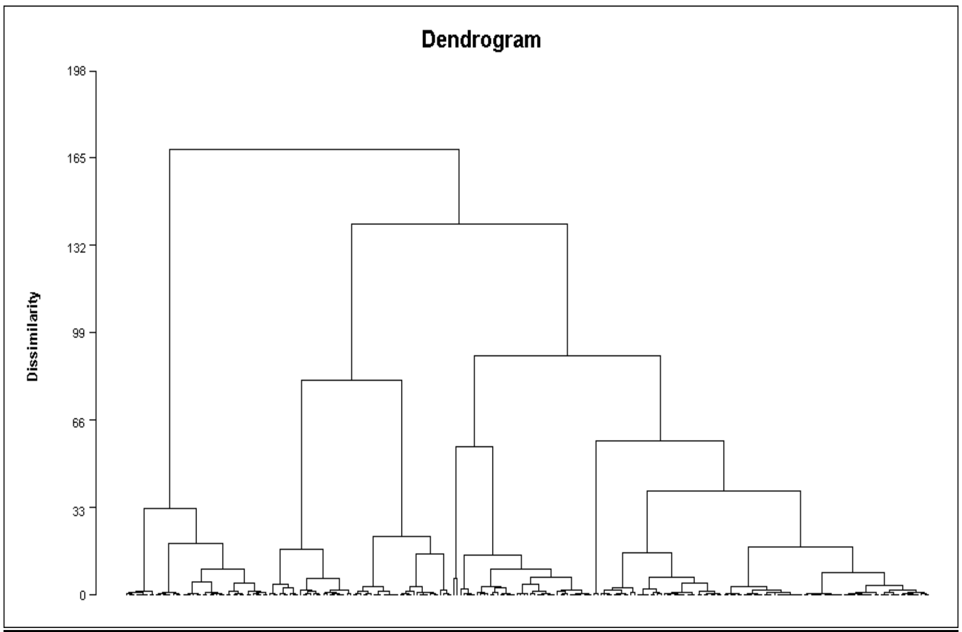
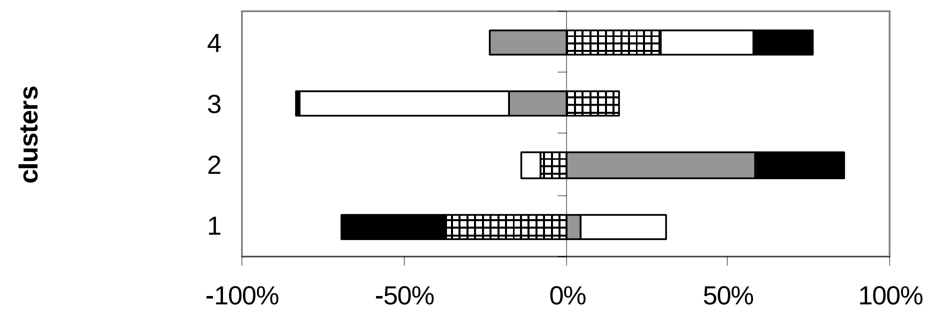


Figure 2 – Dendogram from Ward's Method

The analysis of the previous paragraph constitutes a first approach. However, hierarchical methods impose that once a cluster is formed, it cannot be split. In turn, a non-hierarchical method is more flexible, allowing cases to separate from clusters that they previously integrated. Consequently, following the procedure suggested by several authors (e.g Lattin et al, 2003; Punj and Stewart, 1983), a non-hierarchical k-means clustering procedure has been performed, using the centroids from Ward's method as seeds. Moreover, for the sake of comparing the solution resulting from the present methodology with the four clusters proposed by the European Commission (Table 1) the focus will be mainly on the four-cluster solution.



	1	2	3	4
■ dem fact	-0.42	0.76	-0.03	0.66
□ edu fact	0.35	-0.16	-1.54	1.02
econ fact	0.06	1.59	-0.42	-0.84
(un)emp fact	-0.49	-0.22	0.40	1.05

Figure 3– Final Cluster Centroids

The "fine-tuned" results obtained with the k-means procedure are, at a large extent, coincident with the results of the four-cluster Ward's procedure. More than 86 percent of the regions belong to identical clusters; a number that reaches 100 percent for the cluster that integrates almost all the regions of the new Member States (corresponds to cluster 4 in Figure 3) and is minimum (80 percent) for the largest cluster that integrates the great majority of the EU-15 countries (cluster 1 in Figure 3). The average scores on the four dimensions for the resulting four clusters of NUTS2 European Regions are presented in Figure 3. Significant differences are shown in the profiles of the four clusters: clusters 1 and 2 (particularly this one) show good economic performances; cluster 3 shows an important gap in the education factor; cluster 4 exhibits significant positive values for the unemployment, education and demographic factors.

A more detailed description of the four clusters can be found below and can be further understood by looking at the European map in Figure 4:

Cluster 1 - This is the largest cluster as it is formed by 117 regions mainly situated in **Northern and Central Europe**. Both population and area are around 43 percent of the total. Unemployment is below average as well as is the percentage of active adults versus elderly population. On the other hand, values of the Economic and Education Factors are over the average. Therefore, cluster 1 identifies rich regions, with low unemployment, a wide offering of services, a modern and essentially urban economy and a high percentage of active populations with upper secondary and post-secondary levels. This cluster also integrates two regions from the old Eastern Europe: Közép-Magyarorszá, an Hungarian region that has been classified has a "phasing-in region" by the EC, and, more surprisingly, Estonia, a country that is classified as a Convergence Region due to its low gdp per capita. In this case, a closer analysis revealed that Estonia exhibits other characteristics in terms of unemployment, percentage of active population, percentage of service employment and percentage of active population with tertiary education that differ from the average figures of the other eastern countries and, in particular, from its Baltic neighbors Latvia and Lithuania. So, it can be said that Estonia exhibits signs of being a richer country than, truly, its GDP per head still shows.

Cluster 2 - There are thirty-one regions in this cluster and thirteen of them are some country **capital regions**; for example, Vien (Vienna), Brussels, Berlin, Madrid, Île de France (Paris), Attiki (Athens), Luxemburg, Southern and Eastern Ireland (Dublin), Noord-Holland (Amsterdam), Stockholm or Inner London. Some are new EU country capitals, such as Praga or Bratislavský. The rest of the regions belonging to this cluster are in Belgium (four), Germany (four), Spain (one), Netherlands (five) or United Kingdom (four). This cluster's population is 16 percent of the total, whereas the area is only 3 percent. It is, therefore, the densest cluster. In relation to the factor values, unemployment is a little below the average as well as is education. On the other hand, values of the Economic and Demographic Factors are considerably over the average. Hence, cluster 2 groups are very rich and dense regions.

Cluster 3 - The fifty-one regions in this cluster represent around 21 percent of population and area of the EU total. Geographically all of the regions are in the **South of Europe** with the single exception of ie01 (Border, Midlands and Western Ireland). Unemployment is over the average, whereas the percentage of active adults versus elderly population is around average. On the other hand, values of the Economic and Education Factors are below the average, especially in the second one. The main characteristic of this cluster is the low levels in non-primary education. Therefore, cluster 3 identifies deprived regions, with some unemployment and low levels of upper secondary and post-secondary levels. In addition, many Cluster 3 Regions are Convergence Regions.

Cluster 4 - There are forty-three regions in this cluster representing around 17 percent of the total EU population and area. All of the forty-three regions belong to the non EU-15 countries, with the exception of six German regions belonging to the former Democratic Republic. Cluster 4 regions have the highest levels of unemployment as well as the highest percentage of upper secondary and post-secondary education levels. Moreover, values for the Economic Factor are far below the average, while the percentage of active adults versus elderly population is below average. So, cluster 4, in addition to detecting **Eastern regions**, is also detecting low income regions with high unemployment. All Cluster 4 Regions are Convergence Regions.

Table 6 –Concordance of the Two Classifications

		Comp.Emp. Regions	**Ph. Out Regions**	**Ph. In Regions**	**Conv. Regions**	Total
	1	105	4	4	4	117
Clusters	**2**	30	1	0	0	31
K-means	**3**	16	6	8	21	51
	4	0	2	0	40	42
Total		151	13	12	65	241

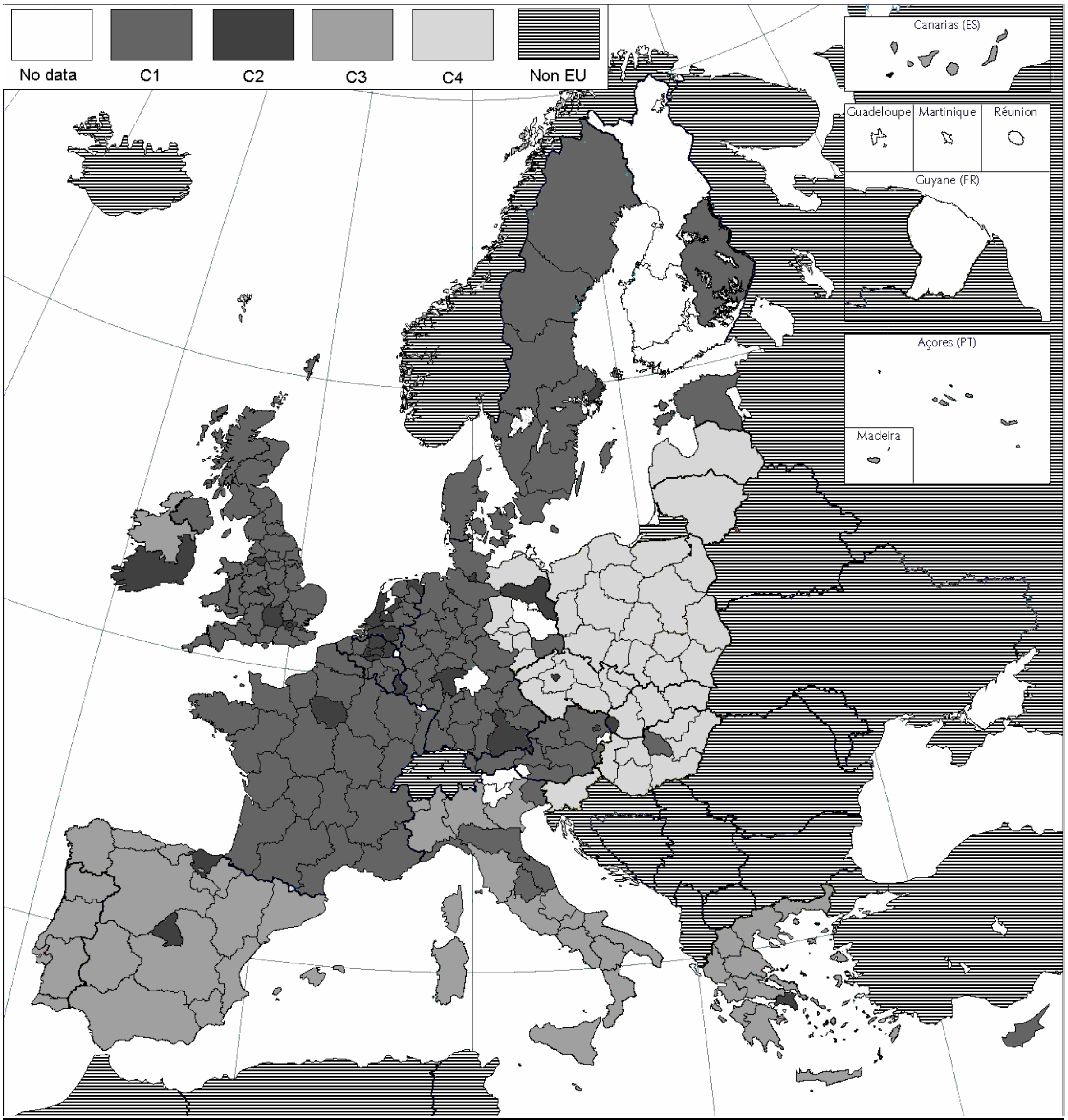


Figure 4 – EU-25: Four Socioeconomic Clusters of Regions (obtained through a non-hierarchical k-means clustering procedure)

The next issue to be analyzed is the degree of concordance between the regional clusters described above and the four classes of regions identified by the European Commission as the basis for European regional policy during the period 2007-2013 (these classes were described earlier in Section 1 and represented in Figure 1). Table 6 shows that clusters 1 and 2 are dominated by *Competitiveness and Employment Regions*; that cluster 4 includes almost exclusively *Convergence Regions*; and, that cluster 3, which includes a great number of southern regions, is spread over all the EC classes. As a conclusion, it is clear that the cluster analysis revealed two different groups within *Competitiveness and Employment Regions* and the same result happened for *Convergence Regions*.

In order to evaluate the significance of this difference and to reach a conclusion about the most convenient classification for pursuing European regional policy targets, it is useful to return to the major regional indicators. Table 7 shows the averages of the different indicators calculated respectively for the EC classification and the four-cluster classification. A comparative analysis reveals that the four-cluster solution allows a more clear distinction among the different regions, which is particularly obvious in cases such as the density of population, GDP per capita and education levels. Notice that the two major groups of the EC classification – *convergence regions* and *competitiveness and employment regions* – include at least two significantly different groups of regions in terms of these indicators. The two other groups - *phasing-in regions* and *phasing-out regions* – spread throughout different clusters in Table 6 and do not have evidently different figures in many indicators. In fact, as GDP per capita, the only basis for the EC classification, varies between 32 percent and 315.4 percent of the European average, the fixing of an exclusive and arbitrary threshold of 75 percent could hardly lead to a better distinction among regions.

Table 7 – Average Indicators for the EC and the Four-Cluster Classifications

Regional Indicators	**Comp.Emp. Regions**	**Conv. Regions**	**Ph.In Regions**	**Ph.Out Regions**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
popdens	481.6	130.6	307.1	190.2	226.3	1544.7	154.6	121.9
pop014	17.0	16.4	16.4	14.7	17.3	17.1	15.0	16.4
pop1564	66.0	68.0	67.3	66.2	65.1	68.3	66.9	69.3
pop65	17.0	15.6	16.3	19.0	17.5	14.5	18.1	14.3
gdppps	114.4	56.7	91.0	78.9	102.6	145.2	87.0	50.7
empagr	3.3	13.9	8.1	7.4	3.5	1.5	12.4	11.9
empserv	68.8	55.9	64.8	64.9	67.7	76.4	59.9	54.6
empind	28.0	30.2	27.2	27.7	28.8	22.0	27.6	33.5
patent	162.1	14.0	30.5	38.1	156.6	201.3	20.8	15.2
emptot	67.3	56.7	60.8	59.5	67.8	67.0	58.3	56.3
empf	60.0	48.3	50.0	49.8	61.3	60.0	44.7	50.8
empm	74.5	65.3	71.7	69.3	74.4	73.9	71.9	61.8
unemptot	6.3	13.5	8.9	11.5	6.2	6.9	9.8	15.3
unemplt	32.7	49.5	36.5	47.4	31.9	34.1	43.1	52.8
unempf	6.7	15.6	11.2	14.5	6.2	7.1	14.2	16.1
unempy	13.6	28.3	19.8	23.5	12.9	14.6	25.1	28.6
lowedu	23.5	26.9	41.0	32.9	19.7	21.3	54.2	11.5
mededu	48.9	55.3	36.4	42.6	53.3	44.4	27.3	70.5
highedu	25.8	17.5	21.8	23.4	24.9	32.1	18.5	17.6

5. Conclusions

Public policies benefit from being based on simple and objective rules, allowing for a transparent implementation by Public Authorities. Sometimes, however, simple and objective rules become established dogmas and should be questioned.

A good example is the deficit limit of three percent of gross domestic product (GDP) established by the EU Stability and Growth Pact. Certainly economic theory and even common good sense can explain why large budget deficits are undesirable and create a burden for future generations. However, there is nothing in economic theory saying that a good limit for deficits is 3 percent and not 2 percent or 4 percent, for instance. In addition, as European Governments have already recognized, the application of the 3 percent rule has to be flexible, taking into consideration what phase of economic cycle a country is facing.

The same reasoning applies to Regional European Policy. The allocation of financial resources has been considerably based on a threshold corresponding to 75 percent of European's average GDP per capita. This rule has already caused the redesign of some NUTS2 regions, in cases where the heterogeneity of the region was negatively affecting the poorest areas (e.g., areas that otherwise would be classified Objective 1, as in Lisboa e Vale do Tejo – Portugal). However, this same rule supports the proposed distribution of funds for the next cohesion period 2007-2013 and the segmentation of European regions shown in Figure 1. In this paper, the authors have shown that this segmentation leads to very heterogeneous groups of regions and, being one-dimensional, is insufficient for characterizing the different domains of dissimilarity among groups, an important issue for designing the application of solutions tailored to the different groups of regions –with their different needs – within the EU territory.

The approach that was followed began by reducing the information of the major regional indicators in four categories – demography, employment, economy and education. The resulting factors were used, with an equal weight, to classify the European regions into four classes for the sake of comparison with the four clusters solution proposed by the European Commission. It was shown that each of the two major groups of the EC classification – *convergence regions* and *competitiveness and employment regions* – comprises at least two significantly different groups of regions, which differ not only in terms of their average income, but also in terms of other indicators. Also, it was revealed that the two other groups - *phasing-in regions* and *phasing-out regions* –, beyond their inexpressive denomination, also seem to lack homogeneity, being spread throughout different clusters.

A final remark: in spite of considering that the statistical techniques that were used in the paper were able to respond to the goals of this research, it seems an interesting and promising task to conduct further analysis aiming to compare results from other different classification techniques.

* In Germany, Brandenburg was divided into two NUTS2 regions. In Spain, Ceuta and Melilla was also divided into two regions. In Italy, the Nord Ovest NUTS1 region was redefined to include Lombardia, previously a NUTS1 region, Nord Est to include Emilia-Romagna, Centro to include Lazio and Sud to include Abruzzo-Moliseand Campania, while a new NUTS1 region, Isole, was formed to cover Sardegna and Sicilia. In Portugal, the former Lisboa e Vale do Tejo NUTS2 region was split between Centro, a new Lisboa region and Alentejo. In Finland, four previous NUTS2 regions in the Manner-Suomi NUTS1 region (all except Itä-Suomi) were reclassified to form three new NUTS2 regions.

** Some authors suggest weighting the eigenvectors by the square roots of their associated eigenvalues, so that the variances of the respective principal components equal the variance accounted for by those components in the original data (Lattin et al., 2003, p. 274). We did not follow this approach since it would lead to different weights for each category.

References

Aragon, Y.; Haughton, D.; Haughton, J.; Leconte, E.; Malin, E.; Ruiz-Gaen, A. and Thomas-Agnan, C. 2003. "Explaining the Pattern of Regional Unemployment: The Case of Midi-Pyrenees Region." *Papers in Regional Science* 82: 155-174.

Barnier, M. 2004. Speech given to the European Parliament's enlarged Conference of Presidents, European Parliament, Brussels, 18 February.

Bartlett, M. S. 1950. "Tests of Significance of Factor Analysis," *British Journal of Psychology* (Statistical Section) 3: 77-85.

Cziráky, D., Sambt, J., Rovan, J. and Puljiz, J. 2005. "Regional development assessment: A structural equation approach," *European Journal of Operational Research*, in press.

European Commission. 2004. "A new partnership for cohesion – convergence, competitiveness, cooperation. Third Report on Economic and Social Cohesion."

Everitt, B. S. 1993. *Cluster Analysis*, 3rd edition., London: Edward Arnold.

Everitt, B. S.; Landau, S. and Leese, M. 2001. *Cluster Analysis*. London: Edward Arnold.

González, J. I. and Morini, S. 2000. "Posicionamiento socioeconómico y empresarial de los municipios de la Isla de Tenerife" (Socioeconomic and business affairs situation in Tenerife Island). *Working Paper 2000-07*. Universidad de La Laguna (in Spanish).

Hair, J. F.; Anderson, R. E.; Tatham, R. L. and Black. 1998. *Multivariate Data Analysis*. Prentice Hall International.

Hill, E. W.; Brennan, J. F. and Wolman, H. L. 1998. "What Is a Central City in the United States? Applying a Statistical Technique for Developing Taxonomies," *Urban Studies* 35,1: 1935-1969.

Kaiser, H. F. 1959. "The Application of Electronic Computers to Factor Analysis," in *Symposium on the Application of Computers to Psychological Problems*, American Psychological Association

Lattin, J., Carrol, J. D. and Green, P. E., 2003. *Analysing Multivariate Data*, Duxbury: Thompson Learning.

Ludlow, L. H. 1999. "The Structure of the Job Responsibilities Scale: A Multimethod Analysis," *Educational and Psychological Measurements* 59,6: 962-975.

Milligan, G. 1980, "An examination of the effect of six types of error perturbation on fifteen clustering algorithms," *Psychometrika* 45 (September): 325-42.

Peschel, K. 1998. "Perspectives of Regional Development around the Baltic Sea," *The Annals of Regional Science* 32: 299-320.

Pettersson, O. (2001): "Microregional Fragmentation in a Swedish County," *Papers in Regional Science* 80: 389-409.

Punj, G. and Stewart, D. 1983. "Cluster Analysis in Marketing Research; a Review and Suggestions for Application," *Journal of Marketing Research* 20 (May): 134-148.

Rovan, J. and Sambt, J. (2003): "Socio-economic Differences among Slovenian Municipalities: A Cluster Analysis Approach," in *Development in Applied Statistics*, Ferligoj. A. and Mrvar A., eds., pp. 265-278.

Rúa Vieytes, A.; Peralta Astudillo, M. J.; Fernández Rodríguez, M. L. and Borrás Pala, F. 2000. *Tipología Socioeconómica de las Regiones Europeas. Comparativa Estadistica* [Socioeconomic Typology of the European Regions. Statistical Comparative], Instituto de Estadística de la Comunidad de Madrid (in Spanish).

Rúa Vieytes, A.; Redondo Palomo, R. and del Campo, C. 2003. "Distribución municipal de la realidad socioeconómica gallega" [Galician Socioeconomic Reality Municipal Distribution], *Revista Galega de Economía* 12,2: 243-262 (in Spanish).

Soares, J. O.; Marquês, M. L. and Monteiro C. F. 2003. "A Multivariate Methodology to Uncover Regional Disparities: A Contribution to Improve European Union and Governmental Decisions." *European Journal of Operational Research* 145: 121-135.

Stimson R.; Baum, S.; Mullins, P. and O'Conor, K. 2001. "A Typology of Community Opportunity and Vulnerability in Metropolitan Australia," *Papers in Regional Science* 80: 45-66.